LCA Methodology

Data Quality

Assessing Input Data Uncertainty in Life Cycle Assessment Inventory Models

[1]Dale J. Kennedy, [2]Douglas C. Montgomery, [3]Dwayne A. Rollier, [4]J. Bert Keats

[1]Ph.D. Candidate, Arizona State University
[2]Professor of Industrial Engineering, Arizona State University
[3]Associate Professor of Industrial Engineering, Arizona State University
[4]Professor of Industrial Engineering, Arizona State University

Corresponding author: Dale J. Kennedy, 5720 W. Park Ave., Chandler, AZ 85226, USA

Abstract

A methodology is presented to develop and analyze vectors of data quality attribute scores. Each data quality vector component represents the quality of the data element for a specific attribute (e.g., age of data). Several methods for aggregating the components of data quality vectors to derive one data quality indicator (DQI) that represents the total quality associated with the input data element are presented with illustrative examples. The methods are compared and it is proven that the measure of central tendency, or arithmetic average, of the data quality vector components as a percentage of the total quality range attainable is an equivalent measure for the aggregate DQI. In addition, the methodology is applied and compared to real-world LCA data pedigree matrices. Finally, a method for aggregating weighted data quality vector attributes is developed and an illustrative example is presented. This methodology provides LCA practitioners with an approach to increase the precision of input data uncertainty assessments by selecting any number of data quality attributes with which to score the LCA inventory model input data. The resultant vector of data quality attributes can then be analyzed to develop one aggregate DQI for each input data element for use in stochastic LCA modeling.

Keywords: Data quality vector; LCA input data quality; LCI input data quality; Life Cycle Assessment; Life Cycle Inventory; stochastic LCA modeling

1 Introduction

The methodology for developing stochastic LCA models presented in KENNEDY, MONTGOMERY, and QUAY (1996) uses a single rating to measure the overall quality of each data element. This rating is based on a sliding scale of one to five, with a one representing the worst quality case, i.e., maximum uncertainty, and a five representing the best quality case, i.e., minimum uncertainty. This same concept is now expanded to enable the LCA practitioner to evaluate individual input data elements using a multitude of quality attributes. The attribute ratings, or scores, for each data element become the components of a data quality vector. These vectors are analyzed to derive aggregate quality ratings for each input data element to support stochastic LCA modeling.

There are cases where aggregated assessments of input data quality are not readily available nor is it advisable in some cases due to poor data quality discrimination resolution. For these instances, an approach similar in concept to the pedigree matrix approach for assessing LCA model input data quality as discussed by WEIDEMA and WESNOES (1995) may be preferable. The method must provide LCA practitioners with a means to state their judgment about the quality of the input data over a limitless array of discrete data quality considerations. Although WEIDEMA and WESNOES stated that the scores in the pedigree matrix are semi-quantitative in that they serve as identification numbers only and should not be aggregated, these values do relate to the overall assessment as presented by FUNTOWICZ and RAVETZ (1990) in the Numeral, Unit, Spread, Assessment, and Pedigree (NUSAP) notation. In the absence of information other than the numeral (i.e., no probability distribution or measure of variance (spread)), either an overall assessment needs to be created to apply the stochastic modeling approach to LCA inventory analyses or, a technique similar to the pedigree matrix that is designed to enable the aggregation of the elements for an overall assessment is needed.

An alternative method to describe the uncertainty in LCA input data elements involves the creation of a $1 \times n$ data quality vector, q. The components of q are established in a similar manner as the single-valued data quality indicator presented in KENNEDY et al. The difference is that the components of q convert LCA practitioner qualitative judgments about specific data quality „attributes" to quantitative indices. Some of the typical attributes LCA practitioners consider were discussed in KENNEDY et al. These included data age, accuracy, completeness, and representativeness of the

total population/process, and frequency of collection/quantity of data collected. Of course, this list of typical attributes is not all inclusive and can vary between LCA practitioners. For example, WEIDEMA and WESNOES include geographical correlation and technological correlation as data quality factors. As the LCA methodology continues to mature and databases are further developed, additional data quality descriptors may be warranted. To accommodate such advances in LCA technology, there is no limit on the number of components, n, in the vector, q.

The DQI development methodology presented here enables the establishment of a data quality vector of any size and an evaluation of the vector to determine the amount of aggregate uncertainty it represents. Analyzing the data quality vector using this methodology results in a single-valued indicator representing the aggregate uncertainty associated with the input data element. The aggregate uncertainty indicator maps directly to the DQI's developed in KENNEDY et al. and reproduced in **Table 1** for ease of reference.

The beta distribution parameters specified in **Table 1** are used to generate random variables for input data in stochastic LCA inventory models. This, in turn, enables making multiple simulation runs of the LCA inventory model to produce results that can be compared using statistical methods.

In the absence of information about the actual input data probability distribution, the beta probability distribution is reasonable to use for several reasons discussed in KENNEDY et al. The beta distribution enables the use of range endpoints and two shape parameters α and β that determine the mean and variance (i.e., spread) of the distribution. As α and β

Table 1: Beta probability distribution parameters for DQI's (baseline, sensitivity level 1, and sensitivity level 2)

Baseline:

	Beta Probability Distribution Parameters	
Data Quality Indicator	Shape Parameters (α,β)	Range Endpoints (±%)
5	5,5	10
4.5	4,4	15
4	3,3	20
3.5	2,2	25
3	1,1	30
2.5	1,1	35
2	1,1	40
1.5	1,1	45
1	1,1	50

Sensitivity Level 1 (SENS L-1):

	Beta Probability Distribution Parameters	
Data Quality Indicator	Shape Parameters (α,β)	Range Endpoints (±%)
5	4,4	20
4.5	3,3	25
4	2,2	30
3.5	1,1	35
3	1,1	40
2.5	1,1	45
2	1,1	50
1.5	1,1	50
1	1,1	50

Sensitivity Level 2 (SENS L-2):

	Beta Probability Distribution Parameters	
Data Quality Indicator	Shape Parameters (α,β)	Range Endpoints (±%)
5	3,3	30
4.5	2,2	35
4	1,1	40
3.5	1,1	45
3	1,1	50
2.5	1,1	50
2	1,1	50
1.5	1,1	50
1	1,1	50

decrease from five to one in **Table 1**, the shape of the distribution becomes flatter indicating higher probability that values closer to the range endpoints will occur for the input data for each run of the stochastic LCA inventory model. Of course, range endpoints resulting from higher percentages of the input data values also indicate greater uncertainty (i.e., lower input data quality) because the input data can assume values over a wider interval.

The sensitivity levels in **Table 1** (i.e., Sensitivity Level 1 (SENS L-1) and Sensitivity Level 2 (SENS L-2)) indicate increasing input data uncertainty as the levels increase. These beta distribution parameters are used when the LCA practitioner is conducting sensitivity analyses to provide the LCA inventory information users an understanding of the sensitivity of the results to under-, and over-, estimation of the input data quality.

The methodology is presented in two sections. The first addresses the development of data quality vectors. The second section presents three data quality vector analysis methods with illustrative examples and a proof of the equivalence of these methods. The aggregation methodology is applied to real-world LCA input data pedigrees that have similar features to the data quality vector. Comparisons of the methods are discussed which leads to the development and analysis of a weighted data quality vector. Finally, some concluding remarks are presented.

2 Methodology

2.1 Data quality vector development

The process of developing data quality vectors begins with the selection of the data quality attributes with which to score the data. This can be done either by a single LCA practitioner or in a group setting. Many group decision techniques, e.g., Nominal Group Technique and the Delphi method (see GOICOECHEA, HANSEN and DUCKSTEIN, 1982), are available to support this part of the process. As LCA techniques continue to mature and more research is accomplished in the study of input data quality, evaluation standards should emerge that standardize the quality attributes to be applied or at least provide a comprehensive set from which LCA practitioners can select.

After the selection of the data quality attributes, the scoring process is similar to that presented in KENNEDY et al. All LCA model input data are individually scored on the same sliding scale for each data quality attribute. The result is a vector of quality indices that represents the practitioner's(s') judgment(s) regarding the uncertainty associated with the data element. The data quality vector is denoted as **q** such that $\{q_i: i = 1,2,...,n\}$ are the set of n data quality attributes that can take on data quality indice values in the range $1 \leq q_i \leq 5$. Note that the scores are not restricted to integer values.

Independence is maintained between assessment scores within the data quality vector. Quality assessment correlation that may exist between individual input data elements, e.g., data having the same age, is not a concern since the overall assessments are accomplished within individual data elements and not across data elements. Correlation between certain input data elements, e.g., chemical stoichiometry and mass balance considerations, does need to be analyzed and accounted for in the stochastic modeling approach, just as it is in the deterministic models.

2.2 Data quality vector analysis for aggregate DQI assignment

To implement the stochastic LCA inventory modeling methodology presented in KENNEDY et al., an aggregate DQI must be derived for each data quality vector for those input data elements with no prior probability distribution information. The aggregate DQI must capture the input data element uncertainty represented by the applicable data quality vector. Several methods are presented to accomplish this analysis. These include linear programming, vector projection, and expected value. Each results in a percentage of maximum attainable quality represented by a given data quality vector.

A proof is presented that provides LCA practitioners the assurance that each of these analysis approaches produce equivalent results. Therefore, LCA practitioners are free to choose the analytical method they find most practical and informative for their particular application.

DQI's are assigned to each LCA input data element according to the percentage of the maximum attainable quality value each represents. The DQI's are assigned as indicated in Table 2. Once the DQI's are assigned to each input data element, the LCA practitioner has the information needed to use **Table 1** to select parameters for the associated beta probability distributions to implement the stochastic LCA modeling methodology. Of course, LCA practitioners should select parameters for the beta, or other applicable input data probability distributions, as appropriate to adequately model any "known" uncertainty regarding individual data elements. For example, if a particular data element should have a skewed distribution so that values closer to one range endpoint are more likely than those near the other, the appropriate beta distribution shape parameters should be chosen independently of **Table 1**.

2.2.1 Linear programming (LP) method

Linear programming (LP) can be used to evaluate the percent of the maximum quality function attained by the data quality vector of interest. WU and COPPIN (1981), HADLEY (1962), and HILLIER and LIEBERMAN (1980) are among many authoritative references for details on LP. The first step in

this process requires the LCA practitioner to formulate the LP problem.

The LP formulation involves the development of the quality objective function to be maximized and the constraints associated with the decision variables. In this application, the decision variables can simply be thought of as the quality indice values that can be assigned to any of the data quality attributes.

The quality objective function is defined as:

$$Q = \sum_{i=1}^{n} q_i \qquad (1)$$

and it is subject to the following constraint space:

$$\left.\begin{array}{l} q_i \le 5 \\ q_i \ge 1 \end{array}\right\} (i = 1,2,\ldots,n)$$

Note that the $q_i \ge 1$ constraint makes the non-negativity constraint required in LP problem formulation redundant.

The solution to the LP can be accomplished by manually applying the simplex method or through graphical methods if the problem is small enough. A number of commercially available software packages with embedded LP problem solvers, such as Microsoft® Excel (see Microsoft® Excel Version 5.0 User's Guide 1994), and Statgraphics® (see Statgraphics® Reference Manual Version 5 1991), can be used to find the maximum of the quality function. The total number of data quality attributes, n, determines the degree of Euclidean (E^n) space that contains the feasible region. Problems of this type in E^2 can be solved using a graphical representation of the LP problem. The LP solution provides the maximum and minimum attainable quality function value (i.e., max Q and min Q).

Of course, for straightforward quality objective functions such as equation (1), the maximum can be readily identified without applying formal LP solution methods. By simple inspection, it can be noted that the maximum attainable quality function value is when all n components have the maximum assignable score. In the case of the one through five sliding scale applied in this paper, where a five is the best assignable score, the maximum attainable unweighted quality function value is represented by $(5n)$.

The next step in the LP method is to determine what percentage of the maximum attainable data quality has been achieved by each data quality vector. The quality function value for each data quality vector must be calculated. The resultant value represents one of an infinite number of parallel iso-quality lines (in E^2) or parallel iso-quality cutting planes (in E^n, n>2). The percent of attainable quality represented by each data quality vector is calculated as follows:

$$\%\ \text{of attainable quality} = \frac{Q - \min Q}{\max Q - \min Q} \times 100 \qquad (2)$$

where min Q is subtracted from Q in the numerator and max Q in the denominator to account for the true percent of attainable quality since min Q>0.

2.2.1.1 Illustrative examples

Consider the case of a data quality vector q containing $n = 2$ data quality attributes. The age of the data might be represented by q_1 and data representativeness by q_2. The formulation of the LP problem becomes (from equation (1)):
Maximize $Q = q_1 + q_2$
subject to:

$q_1 \ge 1$
$q_1 \le 5$
$q_2 \ge 1$
$q_2 \le 5$

The graph of this formulation in E^2 is shown in **Figure 1**. The feasible region is the interior of the square which is bounded by the four constraint equations. Several iso-quality lines have been added to indicate that as the data quality attribute indices improve from worst case (1,1) to best case (5,5), the parallel iso-quality lines continuously improve in value. Q = 10 is maximized at the upper right corner of the feasible region (5,5). As would be expected by simple inspection of the problem formulation, this is the only point that remains feasible to maximize Q. Note that Q = 2 is at a minimum at the lower left corner point (1,1).

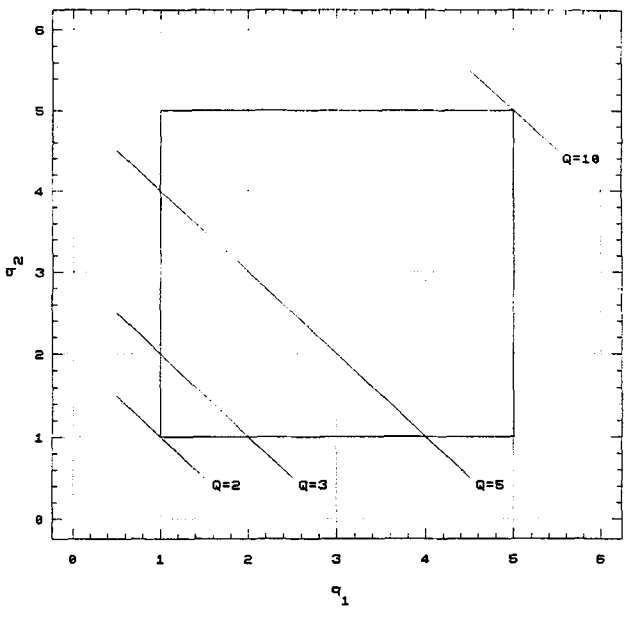

Fig. 1: Feasible region and ISO-quality lines of a two attribute unweighted data quality vector

Evaluating this LP using the Simplex method in Statgraphics® provides the same solution after four pivots. The solution also provides the shadow prices, or margin values, for q_1 and q_2 as 1.0 each. This indicates that quality function value increases a value of 1.0 for each corresponding unit increase in q_1 or q_2. This is increasingly useful information as the complexity of quality objective functions increases.

The percent of the maximum attainable quality (x) for an example data quality vector **q** = (4,2) is determined as follows from equation (1): $Q = 4 + 2 = 6$

From equation (2):

$$x = \% \text{ of attainable quality} = \left(\frac{6-2}{10-2}\right) \times 100$$

$x = 50\%$

Therefore, the aggregate DQI for the (4,2) data quality vector is 3 from **Table 2**. Then, from **Table 1**, stochastic LCA model input data with this aggregate DQI of 3 would be replaced with random variables drawn from a beta probability distribution with range endpoints specified at ±30% of each input datum's value and with shape parameters $\alpha = 1$ and $\beta = 1$. These particular beta distribution shape parameters transform the beta distribution into a uniform probability distribution over the range of possible input data values (i.e., ±30% of the original value in this instance).

Table 3 contains the results of an evaluation of all the combinations of the 1 x 2 integer-valued data quality vectors. Permutations of the vector components are not included because the ordering of vector components in the case of evenly weighted data quality attributes does not alter the result. For instance, the example presented above evaluates the aggregate DQI for the vector **q**=(4,2) as 3 (50%). The evaluation of data quality vector **q**=(2,4) shown in **Table 3** indicates that the same result is obtained.

Table 2: DQI assignment matrix

Achieved Percent (x) of Maximum Attainable Quality Value	Data Quality Indicator (DQI)
$0 \le x < 12.5$	1
$12.5 \le x < 25$	1.5
$25 \le x < 37.5$	2
$37.5 \le x < 50$	2.5
$50 \le x < 62.5$	3
$62.5 \le x < 75$	3.5
$75 \le x < 87.5$	4
$87.5 \le x < 100$	4.5
$x = 100$	5

Table 3: All cases of a 1 x 2 integer-valued data quality vector (max Q = 10; min Q = 2)

Data Quality Vector	Q	% of attainable max Q	Aggregate DQI
(1 1)	2	0	1
(1 2)	3	12.5	1.5
(1 3)	4	25	2
(1 4)	5	37.5	2.5
(1 5)	6	50	3
(2 2)	4	25	2
(2 3)	5	37.5	2.5
(2 4)	6	50	3
(2 5)	7	62.5	3.5
(3 3)	6	50	3
(3 4)	7	62.5	3.5
(3 5)	8	75	4
(4 4)	8	75	4
(4 5)	9	87.5	4.5
(5 5)	10	100	5

A data quality vector containing three attributes is represented in E^3. In this case, the feasible region is the interior of a cube as illustrated in Figure 2. The iso-quality functions are represented as parallel cutting planes. These planes cut through the feasible region and each represents a feasible solution for those data quality vectors contained within the feasible region. All points on the iso-quality plane represent the same quality function value. For evenly weighted data quality attributes, the cutting planes are perpendicular to the optimal data quality vector (5,5,5) just as the iso-quality lines in the two data quality attribute case are perpendicular to the optimal data quality vector (5,5).

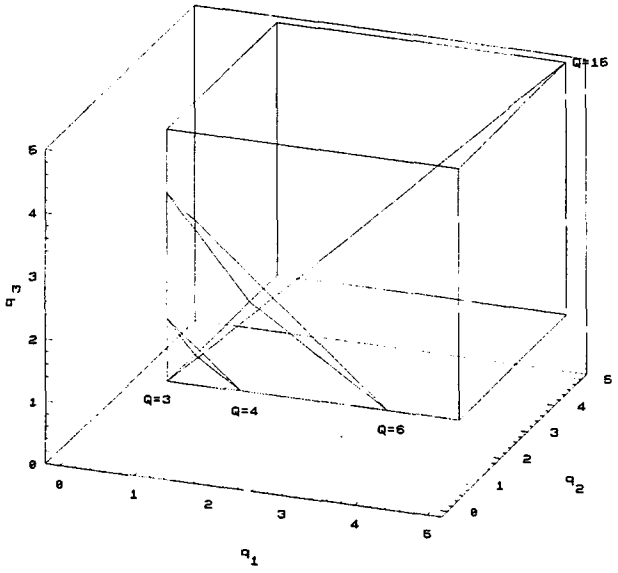

Fig. 2: Feasible region and ISO-quality planes of a three attribute unweighted data quality vector

2.2.2 Vector projection method

An alternate approach is to compute the percent of the magnitude of the optimal data quality vector, **u**, covered by the magnitude of the projection, **p**, of the data quality vector of interest, **q**, on **u**. The percentage of **u** covered by **p** in E^n is defined as follows:

$$\%\ \text{of attainable quality} = \left(1 - \frac{\|\mathbf{u}\| - \|\mathbf{p}\|}{\|\mathbf{u}\|}\right) \times 100 \qquad (3)$$

where PROTTER and MORREY (1964) derive $\|\mathbf{u}\|$ and $\|\mathbf{p}\|$ as:

$$\|\mathbf{u}\| = (\mathbf{u}\cdot\mathbf{u})^{1/2} \text{ and } \|\mathbf{p}\| = (\mathbf{p}\cdot\mathbf{p})^{1/2}$$

Consider the maximum magnitude the $1\times n$ data quality vector, **maxq**, can assume. In this case, $\max q_i = 5$ for $i = 1,2,...,n$. In the same manner, the minimum magnitude the $1\times n$ data quality vector, **minq**, can assume is when $\min q_i = 1$ for $i = 1,2,...,n$. Figure 3 illustrates the E^2 feasible region with the minimum and maximum magnitude vectors, **minq** and **maxq**, displayed. An example data quality vector, **q**=(3,2) is also shown. As indicated in Figure 3, **minq** projects directly on **maxq**. The optimal vector, **u**, is defined as the difference between **maxq** and **minq**. Therefore, **u** is a $1\times n$ vector and $u_i = 4$ for $i = 1,2,...,n$.

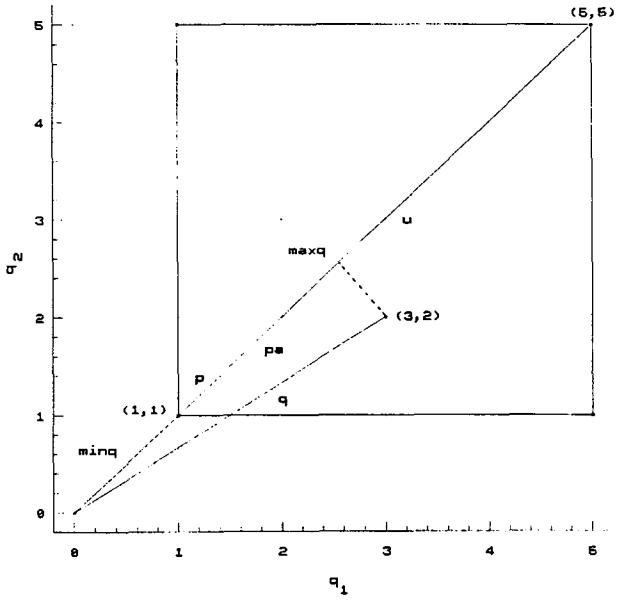

Fig. 3: Vector projection of a two attribute unweighted data quality vector

The projection, **p**, of a vector, **b**, on **a** as derived by FRALEIGH and BEAUREGARD (1987) is $\frac{\mathbf{a}\cdot\mathbf{b}}{\mathbf{a}\cdot\mathbf{a}}\times\mathbf{a}$. Therefore, for any data quality vector, **q**, its projection, **p**, on **u** is $\frac{\mathbf{u}\cdot\mathbf{q}}{\mathbf{u}\cdot\mathbf{u}}\times\mathbf{u}$. Thus,

$$\mathbf{p} = \frac{\sum_{i=1}^{n} u_i q_i}{\sum_{i=1}^{n} u_i^2} \times \mathbf{u}$$

$$= \frac{\sum_{i=1}^{n} u_i q_i}{\sum_{i=1}^{n} 4^2} \times \mathbf{u}$$

$$= \frac{1}{16n} \sum_{i=1}^{n} u_i q_i \times \mathbf{u}$$

Each element of **p** is:

$$p_i = \frac{4}{16n} \sum_{i=1}^{n} 4q_i = \frac{16}{16n} \sum_{i=1}^{n} q_i = \frac{1}{n} \sum_{i=1}^{n} q_i = \bar{q} \quad \text{for } i = 1,2,...,n$$

p must be adjusted by subtracting **minq** so that it is 100% relative to **u**, thus,

$$pa_i = p_i - 1 = \bar{q} - 1 \quad \text{for} \quad i = 1,2,\ldots,n$$

$$\|\mathbf{pa}\| = \sqrt{\sum_{i=1}^{n} (\bar{q} - 1)^2} = \sqrt{n(\bar{q} - 1)^2} = (\bar{q} - 1)\sqrt{n}$$

$$\|\mathbf{u}\| = \sqrt{n4^2} = 4\sqrt{n}$$

From equation (3) and substituting **pa** for **p**:
% of attainable quality

$$= \left(1 - \frac{4\sqrt{n} - (\bar{q} - 1)\sqrt{n}}{4\sqrt{n}}\right) \times 100$$

$$= \left(1 - \frac{[4 - (\bar{q} - 1)]\sqrt{n}}{4\sqrt{n}}\right) \times 100$$

$$= \left(1 - \frac{[4 - \bar{q} + 1]}{4}\right) \times 100 = \left(\frac{\bar{q} - 1}{4}\right) \times 100$$

$$= \frac{1}{4}\left(\frac{1}{n}\sum_{i=1}^{n} q_i - 1\right) \times 100$$

$$= \left(\frac{1}{4n}\sum_{i=1}^{n} q_i - \frac{1}{4}\right) \times 100 \qquad (4)$$

2.2.2.1 Illustrative example

Consider the example data quality vector, **q** = (3,2), illustrated in **Figure 3**. The percent of attainable quality achieved is:

$$\%\ \text{of attainable quality} = \left[\frac{(3+2)}{4(2)} - \frac{1}{4}\right] \times 100 = 37.5\,\%$$

This represents an aggregate DQI of 2.5 from **Table 2**. Note that this is the same result as presented in **Table 3** for the data quality vector **q** = (2,3).

2.2.3 Expected value method

The first step in this method is to find the expected value, $\bar{q}$, of the vector's components. Then, the percent of attainable quality is determined by computing the percent of the range of aggregate data quality indicators, i.e., max q_i-min q_i, that $\bar{q}$ represents. Since the range of values the q_i can assume is $1 \le q_i \le 5$, the quality range is 4. Also, similar to the LP and vector projection methods, $\bar{q}$ must be adjusted by subtracting min q_i to account for the true percent of quality range since min $q_i > 0$.

$$E(Q) = \bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i \qquad (5)$$

From equation (5):

% of attainable quality

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} q_i - 1}{range} \times 100 = \frac{\frac{1}{n}\sum_{i=1}^{n} q_i - 1}{4} \times 100$$

$$= \frac{\frac{1}{n}\sum_{i=1}^{n} q_i - 1}{4} \times \left(\frac{n}{n}\right) \times 100$$

$$= \frac{\frac{n}{n}\sum_{i=1}^{n} q_i - n}{4n} \times 100$$

$$= \left(\frac{1}{4n}\left[\sum_{i=1}^{n} q_i - n\right]\right) \times 100$$

$$= \left(\frac{1}{4n}\sum_{i=1}^{n} q_i - \frac{1}{4}\right) \times 100 \qquad (6)$$

2.2.3.1 Illustrative example

Consider the $n = 3$ component example of a data quality vector, **q** =(2.4,5,4.1). The percent of attainable quality achieved is from equation (6):

x = % of attainable quality

$$= \left(\left\{\frac{1}{4(3)}[2.4 + 5 + 4.1]\right\} - \frac{1}{4}\right) \times 100$$

x = 70.83%

This represents an aggregate DQI of 3.5 from **Table 2**.

2.3 Proof of equivalence of analysis methods

It is obvious from the development of the vector projection and expected value analysis methods that both result in the same aggregate value (reference equations (4) and (6)). However, the equivalence of the LP analysis method is not readily apparent without further development. The basic structure of the LP formulation enables the use of an equivalent numerical approach to determine the aggregate DQI from equation (2) as follows:

% of attainable quality

$$= \frac{\sum_{i=1}^{n} q_i - \sum_{i=1}^{n} 1}{\sum_{i=1}^{n} 5 - \sum_{i=1}^{n} 1} \times 100$$

$$= \frac{\sum_{i=1}^{n} q_i - n}{5n - n} \times 100$$

$$= \left(\frac{1}{4n} \left[\sum_{i=1}^{n} q_i - n \right] \right) \times 100$$

$$= \left(\frac{1}{4n} \sum_{i=1}^{n} q_i - \frac{1}{4} \right) \times 100 \qquad (7)$$

Equations (4), (6), and (7) are all identical, proving that any of these data quality vector analysis methods result in the same DQI assignment. Note that these are for the rating scale of one to five with five being the best value. The analysis methods are easily adapted to any rating scale the LCA practitioner prefers.

3 Application of Analysis Methodology to Real-World LCA Data Pedigrees

This method is also applied to the pedigrees presented in WEIDEMA and WESNOES. The authors present 27 pedigrees for the data contributing to the energy consumption for crop production in the unallocated life cycle of rye bread. They also present the basic coefficient of variation (CV) for each datum and a modified CV (and mean, μ) for each datum after incorporating additional uncertainties relating to the datum's pedigree.

Table 4 presents a subset of this data along with the comparable results that would be attained by applying the data quality vector analysis methodology to the pedigrees. The pedigrees were established based on the same scoring scale of one to five, however the scale is reversed. A one represents the best case and a five the worst. Therefore, the pedigree scores presented in Table 4 are transformed for applicability to the aggregation methodology presented here. For example, a one becomes a five and vice versa, a two becomes a four and vice versa, and a three remains a three. Also, an additional column has been added to the WEIDEMA and WESNOES information by applying a ±3s to the mean based on $CV = \frac{s}{\bar{x}}$ where $s = \hat{\sigma}$. Assuming a normal distribution applies, this is meant to provide an indication of the percentage of the mean that captures the majority (99.7%) of the values the input data may assume. A similar indicator from the beta probability distribution parameterization presented in KENNEDY et al. and reproduced in Table 1 is applied to the aggregate DQI for all sensitivity levels.

Table 4: Application of data quality vector analysis methodology to the data presented in WEIDEMA and WESNOES (1995) using transformed pedigree matrices

WEIDEMA and WESNOES (1995) Data (Note: the ±3s column is derived from the modified CV column to facilitate comparison)					Results of Applying the Data Quality Vector Analysis Methodology to the "Transformed" Data Quality Index			
Data	Basic CV (%)	Data Quality Index	Modified CV (%) (and μ)	±3s [a]	Aggregate DQI	Baseline Case (α,β) [b]	SENS L-1 (α,β) [b]	SENS L-2 (α,β) [b]
1	1	(4 5 5 5 4)	10	$\pm.30\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (4,4)	$\pm.35\bar{x}$ (2,2)
2	1	(5 5 5 5 5)	1	$\pm.03\bar{x}$	5	$\pm.10\bar{x}$ (5,5)	$\pm.20\bar{x}$ (4,4)	$\pm.30\bar{x}$ (3,3)
3	19	(5 5 5 4 5)	25	$\pm.75\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
4	8	(5 5 5 4 5)	12	$\pm.36\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
5	9	(5 5 5 4 5)	13	$\pm.39\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
6	6	(5 5 5 4 5)	18	$\pm.54\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
7	10	(5 5 5 4 5)	20	$\pm.60\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
8	--	(5 5 5 3 5)	25	$\pm.75\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
9	14	(5 5 5 4 5)	22	$\pm.66\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
10	11	(5 5 5 4 5)	20	$\pm.60\bar{x}$	4.5	$\pm.15\bar{x}$ (4,4)	$\pm.25\bar{x}$ (3,3)	$\pm.35\bar{x}$ (2,2)
11	6	(4 5 5 4 2)	37	$\pm1.11\bar{x}$	4	$\pm.20\bar{x}$ (3,3)	$\pm.30\bar{x}$ (2,2)	$\pm.40\bar{x}$ (1,1)
12	6	(4 5 5 1 1)	71	$\pm2.13\bar{x}$	3	$\pm.30\bar{x}$ (1,1)	$\pm.40\bar{x}$ (1,1)	$\pm.50\bar{x}$ (1,1)
13	59	(3 5 1 3 2)	62 (μ 20% higher)	$\pm1.86\bar{x}$	2.5	$\pm.35\bar{x}$ (1,1)	$\pm.45\bar{x}$ (1,1)	$\pm.50\bar{x}$ (1,1)

[a] Assuming the applicable probability distribution is the normal, ±3s represents 99.7% of the values that the data element can assume as a random variable

[b] The (α,β) shape parameters provide additional information about the likelihood of drawing random variable data values from the extremes of the range on $\bar{x}$ (see KENNEDY, MONTGOMERY and QUAY (1996)) In this case, the ±% $\bar{x}$ represents 100% of the values the data element can assume as a random variable regardless of the distribution parameters

According to FUNTOWICZ and RAVETZ (1990), the application of aggregation methodology to the pedigrees is acceptable given no other available information about the NUSAP assessment of the data. However, it is important to note that a direct one-to-one comparison in methods is not possible because the approaches differ somewhat. The intent is to demonstrate the usefulness of the aggregation methodology to real-world data and to discuss those comparisons that are relevant.

The WEIDEMA and WESNOES modified CV column indicates that in instances of low input data quality, the input data random variables can assume negative values. The authors do not address this circumstance. Obviously, control boundaries (i.e., range endpoints) would need to be applied in the modeling approach to prevent the use of negative values. The data quality vector methodology results in range endpoint selections that ensure the input data random variables can assume only positive values.

Many of the pedigrees in Table 4 are the same, e.g., data elements three through seven, nine, and ten have the same data quality index. It would seem that the same pedigree, or data quality index, should represent the same degree of uncertainty about the data element. This is the case with the data quality vector methodology. However, this is not the case with the WEIDEMA and WESNOES approach. Although the authors do not address exactly why this occurs, it appears that additional weighting or fractional data quality assessments are being made using the pedigree as a guide. The data quality vector methodology can account for such fractional assessments within the vector. It can also accommodate data quality attribute weighting with some minor adjustments to the aggregate quality function formulations.

4 Data Quality Vectors with Weighted Data Quality Attributes

In most cases, the input data quality attributes will be evenly weighted. However, there will undoubtedly be instances when the LCA practitioner(s) will want to weight one quality attribute more heavily than another. For example, the energy consumption requirements of the processing technology associated with the bottling system alternative LCA models may be changing quickly over time. New or modified processing methods may be altering processing efficiencies. In this instance, the LCA practitioner may want to weight the "age of data" quality attribute more heavily than the other quality attributes to account for such volatility. In the WEIDEMA and WESNOES rye bread LCA example, one of the data quality goals was that recent data was preferred over other data quality aspects (i.e., attributes). The weighted data quality attribute method can be used to express this preference so that the resultant aggregate DQI reflects the appropriate additional or reduced data uncertainty. In this way, independence is maintained between data quality attribute judgments.

As in assigning the quality scores for each attribute, LCA practitioners must also select the magnitude of the data quality attribute weights using judgment. There are a number of techniques available to assist with this process. One of the more extensive groups of decision analysis tools available are those associated with multiattribute utility theory (MAUT). GOICOECHEA, HANSEN and DUCKSTEIN (1982) provide extensive detail on the MAUT methods as well as an assortment of less formal group decision making techniques.

Incorporating weighted data quality attributes is a straightforward extension of the data quality vector method. Consider an associated weight vector **w** such that $\{ w_i : i = 1,2,...n \}$ are the set of n data quality attribute weights. The quality function objective is now expressed as:

$$Q = \sum_{i=1}^{n} w_i q_i \qquad (8)$$

subject to (as before):

$$\left.\begin{array}{l} q_i \le 5 \\ q_i \ge 1 \end{array}\right\} (i = 1,2,\ldots,n)$$

No special conditions on **w** are necessary. The w_i, like the q_i, are real or integer values. The data quality attribute weight vector is normally a unit vector and, as such, has no effect on the quality function. There is no reason to select a weight of zero for a particular attribute because the effect would be to remove that attribute and, thus, its quality indice, from further consideration. This is better accomplished by evaluating the applicable n - 1 data quality vector.

The effect of the w_i is limited to the objective function. Any data quality attribute weight vector, other than the unit vector, will change the slope of the objective quality function. It has no effect on the feasible region as indicated by the constraint set.

As before, the quality function is evaluated for individual weighted data quality vectors. The result is divided by the maximum attainable value for the weighted quality function found using LP. This provides the same relative percentage as before so that the same aggregate DQI assignment policy is applicable. This, in turn, enables use of the same beta probability distribution parameterization as before.

4.1 Illustrative example

Again, consider the case of a data quality vector **q** containing $n = 2$ data quality attributes. The associated weight vector **w** is (2, 0.5). The formulation of the LP problem using equation (8) becomes:

Maximize $Q = w_1 q_1 + w_2 q_2 = 2q_1 + 0.5q_2$
subject to:

$q_1 \ge 1$
$q_1 \le 5$
$q_2 \ge 1$
$q_2 \le 5$

The graph of this formulation in E^2 is shown in Figure 4. The feasible region is the same as was indicated in Figure 1. Several iso-quality lines have again been added to show the change in slope that occurs from the weighted quality objective function variables. The iso-quality lines are no longer perpendicular to the optimal vector (5,5) which indicates that the quality value can be increased a greater amount per unit of q_1 or q_2 by increasing the heavier weighted quality vector component. In this case, the upper right hand corner of the feasible region, i.e., (5,5) indicates that Q = 12.5 is the maximum. The minimum is at (1,1) where Q =2.5.

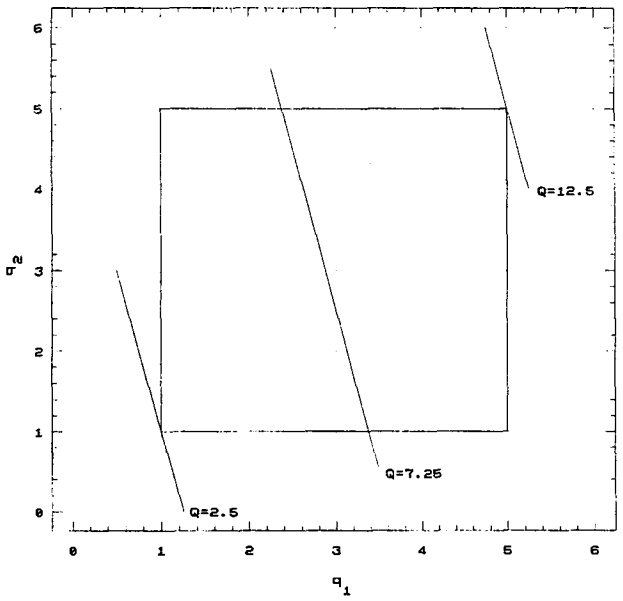

Fig. 4: Feasible region and ISO-quality lines of a two attribute unweighted data quality vector (w = (2, 0.5))

Evaluating this LP using the Simplex method in Statgraphics® provides the same solution after four pivots. The solution also provides the shadow prices, or margin values, for q_1 and q_2 as 2.0 and 0.5 respectively. This indicates that the quality function value increases a value of 2.0 for each corresponding unit increase in q_1 and a value of 0.5 for each corresponding unit increase in q_2.

The percent of the maximum attainable range of Q for an example data quality vector q = (4.5, 2.5) is from equation (8): Q = 2(4.5) + 0.5(2.5) = 10.25
From equation (2):

$$x = \%\text{ of attainable quality} = \left(\frac{10.25-2.5}{12.5-2.5}\right)\times 100$$

$$x = 77.5\%$$

Therefore, the aggregate DQI for the (4.5, 2.5) data quality vector is 4 from Table 2.

4.2 Proof of equivalent analysis methods

As before, the ratio of the expected value of the weighted data quality vector to the maximum attainable weighted data quality range is equivalent to the LP analysis method. In this case, the weighted mean, $\bar{q}_w$, must be determined and used in the percent of total quality range computation as above. HAMBURG (1983) defines the weighted mean as:

$$\bar{q}_w = \frac{\sum_{i=1}^{n} w_i q_i}{\sum_{i=1}^{n} w_i} \tag{9}$$

Therefore, from equation (9):
% of attainable quality

$$= \frac{\frac{\sum_{i=1}^{n} w_i q_i}{\sum_{i=1}^{n} w_i} - 1}{range}\times 100 = \frac{\frac{\sum_{i=1}^{n} w_i q_i}{\sum_{i=1}^{n} w_i} - 1}{4}\times 100$$

$$= \left(\frac{\sum_{i=1}^{n} w_i q_i}{4\sum_{i=1}^{n} w_i} - \frac{1}{4}\right)\times\left(\frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} w_i}\right)\times 100$$

$$= \frac{\sum_{i=1}^{n} w_i q_i}{4\sum_{i=1}^{n} w_i} - \frac{\sum_{i=1}^{n} w_i}{4\sum_{i=1}^{n} w_i}\times 100$$

$$= \frac{\sum_{i=1}^{n} w_i q_i - \sum_{i=1}^{n} w_i}{4\sum_{i=1}^{n} w_i}\times 100 \tag{10}$$

The LP approach requires further development to complete the proof. From equations (2) and (8):
% of attainable quality

$$= \frac{\sum_{i=1}^{n} w_i q_i - \sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} 5w_i - \sum_{i=1}^{n} w_i}\times 100$$

$$= \frac{\sum_{i=1}^{n} w_i q_i - \sum_{i=1}^{n} w_i}{5\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} w_i}\times 100$$

$$= \frac{\sum_{i=1}^{n} w_i q_i - \sum_{i=1}^{n} w_i}{4\sum_{i=1}^{n} w_i}\times 100 \tag{11}$$

Note that equations (10) and (11) are identical. Also, as before, this proof is based on the one to five sliding scale with five representing the best case.

5 Conclusions

The DQI development extension methodology presented does not require complex mathematical analyses. The increased precision (i.e., fractional estimates) in the data quality indices provided for the LCA models presented in KENNEDY et al., are directly applicable and do not contribute to increased cost or loss of effectiveness. To the contrary, for those practitioners inclined to use fractional estimates as a means to quantify further resolution in the uncertainty of the data, the increased precision results in a more fine-tuned uncertainty analysis. As yet, the effect on the precision of individual input data elements is difficult to discern. With thousands of input variables, it is a difficult task to evaluate the main effects and interaction effects up to the nth order. Therefore, LCA practitioners should be encouraged to provide as much precision in the DQI's as they feel is reasonable to assess.

FUNTOWICZ and RAVETZ report that in their experience, there is a remarkable degree of agreement on the pedigree ratings among experts within the general area of competence. This is attributed to the fact that once the modes are well defined and understood by the evaluators, there is very little room for disagreement. Therefore, the pedigree is considered to be a robust indicator of the strength of the associated NUSAP assessment. Likewise, the LCA practitioners that assigned the DQI's to the data in the models presented in KENNEDY et al. were in general agreement on the approach and the outcome once an understanding of the sliding scale was achieved among the raters. It is expected that the robust nature of the pedigree and single-valued DQI's will also be characteristic of the data quality vector.

As an alternative to the weighted method, LCA practitioner(s) might prefer to establish more detailed constructs for the individual data quality attribute indices in an attempt to account for preferences between attributes. However, the weighted data quality attribute approach is more efficient and effective. The decision on what weights to use is accomplished once and does not require stronger quality attribute constructs that may be difficult to develop such that the entire data set is adequately represented. In addition, the weighted data quality attribute approach guarantees equivalence is maintained across all LCA model input data for all associated attributes. Whereas, the individual scoring of each data element, even using a more detailed scoring construct, requires the quantification of additional LCA practitioner subjectivity across all data elements. This naturally leads to a reduction in the equivalence of quality assessments between data.

The main benefit to evaluating the problem using the linear programming approach is the ability to readily determine, as in the weighted case, which quality component(s) to improve to minimize uncertainty (maximize quality) in any particular data element. This is done by analyzing the shadow prices (quality margin values). Since the quality function is additive, most of this analysis can be done by inspection of the quality function once the LCA practitioner gains a thorough understanding of the concept and becomes skilled at interpreting these models. The LP approach is also useful for evaluating any special cases of data quality vectors that may generate the need to develop and evaluate more complex quality objective functions.

The LCA practitioner can be assured that regardless of the spread indicated by the data quality vector components, the ratio of the arithmetic average of the vector components to the total quality range attainable provides the appropriate measure of aggregate quality. This also applies in the weighted average case. The aggregation methods presented are also easily adapted to other numerical rating scales the LCA practitioner may prefer. The resultant aggregate DQI's map directly to the DQI's used for stochastic LCA modeling.

6 References

FRALEIGH, J.B.; BEAUREGARD, R.A. (1987): Linear Algebra. Addison-Wesley Publishing Company, Inc., MA

FUNTOWICZ, S.O.; RAVETZ, J.R. (1990): Uncertainty and Quality in Science for Policy. Kluwer Academic Publishers, The Netherlands

GOICOECHEA, A.; HANSEN, D.R.; DUCKSTEIN, L. (1982): Multiobjective Decision Analysis with Engineering and Business Applications. John Wiley & Sons, Inc., NY

HADLEY, G. (1962): Linear Programming. Addison-Wesley Publishing Company, Inc., MA

HAMBURG, M. (1983): Statistical Analysis for Decision Making. 3rd Ed. Harcourt Brace Janovich, Inc., NY

HILLIER, F.S.; LIEBERMAN, G.J. (1980): Introduction to Operations Research. 3rd Ed. Holden-Day, Inc., CA

KENNEDY, D.J.; MONTGOMERY, D.C.; QUAY, B.H. (1996): Stochastic Environmental Life Cycle Assessment Modeling: A Probabilistic Approach to Incorporating Variable Input Data Quality. Int. J. LCA 1 (4)

Microsoft® Excel Version 5.0 User's Guide (1994): Microsoft Corporation, Kirkland, WA

PROTTER, M.H.; MORREY, C.B. Jr. (1964): Modern Mathematical Analysis. Addison-Wesley Publishing Company, Inc., MA

Statgraphics® Reference Manual Version 5 (1991): STSC, Inc., Rockville, MD

WEIDEMA, B.P.; WESNOES, M.S. (1995): Data Quality Management for Life Cycle Inventories – An Example of Using Data Quality Indicators. Presented to the 2nd SETAC World Congress, Vancouver

WU, N.; COPPIN, R. (1981): Linear Programming and Extensions. McGraw-Hill Book Company, NY